REMARKS

Claims 1-79 were pending in the application. Claims 23, 24, 30-66, and 73-79 were withdrawn from consideration as directed to non-elected inventions.

Claims 1, 3, 12, 25-27, and 67 have been amended. New claim 80 has been added. Support for the amendments can be found throughout the application as originally filed.

Claim 22 has been canceled without prejudice to its presentation in future, related applications.

The title has been replaced.

Upon entry of this amendment claims 1-21, 25-29, 67-72, and 80 will be pending.

No new matter has been added.

Information Disclosure Statement

The Office alleges that the references listed on the PTO-1449, (filed September 18, 2001 were not present in the current application file. Copies of the PTO-1449 filed on September 18, 2001, and the references cited therein that were apparently misplaced by the USPTO will be sent under separate cover. Applicant notes that the Examiner will "consider them as though they were submitted with the IDS in paper No. 4" (Office Action, page 2).

Priority

The Office alleges that the present claims are not supported in the manner required by 35 U.S.C. § 101 and 112, first paragraph, by the priority application and, therefore, that the present claims are not entitled to the benefit of the filing date of the priority application. The Office alleges that the priority application fails to provide any specific, substantial and credible utility and provides no guidance or working examples to teach how to use the claimed invention. Applicant respectfully disagrees.

The crux of the Office's rejection of the priority claim is similar to the rejections set forth in the present application, in that the pending claims allegedly lack utility and

are not enabled. However, as discussed below, the pending claims have utility and enable a person of skill in the art to make and/or use the claimed invention. Since the prior application's disclosure is similar to the present application (see, for example, pages 34-48 of Provisional Serial No. 60/225,262), when the pending claims are found to have utility and be enabled, the prior application must also satisfy the requirements under 35 U.S.C. § 101 and 112, first paragraph. Therefore, Applicant respectfully requests that the effective filing date of the present application be recognized as the filing date of the priority application, August 15, 2000.

Title

The Office has objected to the title as not being descriptive. Although Applicant disagrees, in order to further prosecution, Applicant has replaced the title with an even more descriptive title.

Objections

The specification stands objected to as allegedly failing to provide proper antecedent basis for the claimed subject matter. Specifically the Office alleges that it is unable to find basis in the specification for the limitation in claim 72 that "a host cell according to claim 71 that has been *co-transfected* with a polypeptide [sic]." (emphasis in original, Office Action, page 3). Applicant respectfully disagrees.

As an initial matter, it is unclear what part of claim 72 the Office objects to. It appears that the Office objects to the use of the term "co-transfect" since the Office italicized the term. If this is incorrect, however, Applicant respectfully requests that the Office further clarify this objection.

As an initial matter, Applicant respectfully points out that the claim recited in the Office Action is not an accurate quotation of claim 72 as filed. Claim 72 as filed recites:

> A host cell according to claim 71 that has been co-transfected with a ***polynucleotide*** encoding the nGPCR-1079 amino acid sequence set forth in a sequence of SEQ ID NO:1 and that expresses the nGPCR-1079 having the amino acid sequence set forth in SEQ ID NO:2.

(emphasis added). Therefore, if the Office was objecting to the recitation of "co-transfected with a ***polypeptide***," (emphasis added) Applicant respectfully requests that this objection be withdrawn in view of the correct quotation of claim 72.

However, if the objection to claim 72 is based on the term "co-transfected" Applicant respectfully disagrees that there is no "basis" in the specification for the term. Claim 72, as set forth above, was filed "as is" with the present application. Therefore, even if there were no explicit mention of the term "co-transfected" in the remainder of the specification, the disclosure in the claim itself would serve as written description support as support for the claim language can be found in the claim itself. (see, M.P.E.P. §2163.02.) Further, the term "co-transfected" is well known to one of ordinary skill in the art and, in reference to claim 72, refers to at least two nucleotide sequences being transfected together into a host cell. The term "co-transfected" also appears in the specification and is therefore clearly supported by the present application. (see, for example, page 78, paragraph [00282], and page 79, paragraph [00286]).

In view of the foregoing, Applicant respectfully requests that the objection to claim 72 be withdrawn.

Claim 67 stands objected to for being dependent upon a non-elected claim. Claim 67 has been amended so that is no longer depends on a non-elected claim, rendering this objection moot. In view of the foregoing, Applicant respectfully requests that the objection to claim 67 be withdrawn.

Rejection under 35 U.S.C. § 101

Claims 1-22, 25-29, and 67-72 stand rejected under 35 U.S.C. § 101 because the claimed invention is allegedly not supported by a specific, substantial and credible asserted utility or a well established utility. The Office also alleges that the asserted utilities are "not considered specific, or substantial because the specification fails to provide specific support for these uses, nor any information about the ligand, a particular function, or biological significance of the polypeptide encoded by the nucleic acid." (Office Action, page 4). Applicant respectfully disagrees.

The Claimed Invention Has A Specific Utility

To meet the utility requirement, the invention must be "practically useful," *Anderson v Natta*, 480 F.2d 1392, 1397 (CCPA 1973) and confer a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534 (1966). The threshold of utility under this standard is not high, and requires merely an "identifiable" benefit. *Juicy Whip Inc.v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999). In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180 (Fed. Cir. 1991), the CAFC explained that "An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility." *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

Inventions that achieve a practical use, a use that is also achieved by other inventions, satisfy the utility requirement. Thus practical utilities can be directed to classes of inventions, so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. *Montedison*, 664 F.2d at 374-75. For example, many materials conduct electricity. This general utility applies to a broad class of inventions (conductive materials) and satisfies the utility requirement of section 101. The fact that other materials also conduct electricity does ***not*** mean that other materials that conduct electricity want for utility. What is important, however, is that G protein-coupled receptors (GPCRs) are known to have practical uses well beyond throwaway uses like snake food.

Practical uses for GPCRs include therapeutic and diagnostic uses as well as research-based uses. Many medically significant biological processes are mediated by signal transduction pathways involving G-proteins and other second messengers, and GPCRs are recognized as important therapeutic targets for a wide range of diseases. According to a recently issued United States patent, nearly 350 therapeutic agents targeting GPCRs have been successfully introduced onto the market in only the last fifteen years. (See U.S. Patent No. 6,114,127, at col. 2, lines 45-50.) A recent journal

review reported that most GPCR ligands are small and can be mimicked or blocked with synthetic analogues. That, together with the knowledge that numerous GPCRs are targets of important drugs in use today, make identification of GPCRs "a task of prime importance." (See, Marchese et al., Trends Pharmacol. Sci., 20(9): 370-5, 1999, attached hereto). Thus, the allegation that there is no well established utility for proteins of the class that the Applicant is now claiming is directly refuted by industry evidence.

The Office appears to be under the impression that inventions that are, *inter alia*, useful for use in research, are unpatentable. This is not true. The Patent Office's patent database is replete with patents claiming useful research tools, *e.g.*, spectrophotometers. A material whose only use is as a tool in research may indeed be patentable. *Brenner* excludes only those research purposes where the ***only*** use of the material itself is as the subject of research. If *Brenner* had held otherwise, any chemical material would, by virtue of its existence, be useful. However, nowhere do those cases state or imply that a material cannot be patentable if has some other beneficial use in research.

Assay methods, like many other tools used in research, have an immediately realizable "real world" value. For example, an assay method that can identify chemical compounds that possess a particular physical, structural or biological property clearly has "real world" value irrespective and independent from the utility that may be associated with the compounds identified using the assay method. As a consequence, a presumption that assay methods cannot possess utility if the compound isolated or identified using the assay do not have utility would be the product of a flawed analysis of *Brenner*. Such a conclusion also would suggest that processes and products can never possess utility if their utility lies in the field of research. Indeed, the application of this concept of the utility requirement as it relates to methods for assaying or identifying compounds, if taken literally, would mean that claims to methods such as NMR, infrared, x-ray crystallography, and screening for other important biological properties, would be unpatentable because further research would be necessary to establish utility for the compounds identified or assayed. This certainly cannot be the result intended by the Patent Office when issuing the Utility Examination Guidelines.

Genes encoding GPCRs can also be used, for example, for toxicology testing to generate information useful in activities such as drug development, even in cases where little is known as to how a particular GPCR works. No additional experimentation would be required, therefore, to determine whether a GPCR has a practical use as all GPCRs have at least one practical use.

Because all GPCRs, as a class, convey practical benefit (much like the class of DNA ligases identified in the Training Materials), there should be no need to provide additional information about them. A person of ordinary skill in the art need not guess whether any given GPCR conveys a practical benefit. Nor is it necessary to know how or why any given GPCR works. It is settled law that how or why any invention works is irrelevant to determining utility under 35 U.S.C. §101: "[I]t is not a requirement of patentability that an inventor correctly set forth, or even know, how or why the invention works." *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999)(quoting *Newman v. Quigg*, 877 F.2d 1575, 1581 (Fed. Cir. 1989).

Applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner*, 383 U.S. at 532. The amount of evidence required to prove utility depends on the facts of each particular case. *In re Jolles*, 628 F.2d 1322, 1326 (CCPA 1980). "The character and amount of evidence may vary, depending on whether the alleged utility appears to accord with or to contravene established scientific principles and beliefs." *Id.* Unless there is proof of "total incapacity," or there is a "complete absence of data" to support the applicant's assertion of utility, the utility requirement is met. *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 (Fed. Cir. 1992); *Envirotech*, 730 F.2d at 762. The Office has failed to provide proof of "total incapacity", and Applicant has provided information that supports the asserted utilities.

The Office is also reminded that a patent applicant's assertion of utility in the disclosure is presumed to be true and correct. *In re Cortwright*, 165 F.3d at 1356; *Brana*, 51 F.3d at 1566. If such an assertion is made, the Patent Office bears the burden to demonstrate that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved. *Id.* To do so, the PTO must provide evidence or sound

scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566.

Applicant has demonstrated a "substantial likelihood" of utility by showing a "reasonable correlation" between the utility of the known composition and the composition being claimed. *Fujikawa v. Wattanasin*, 93 F.3d 1559, 1565 (Fed. Cir. 1996). The presently claimed GPCR is related to known GPCRs. The Office has not provided evidence or sound scientific reasoning that one skilled in the art would doubt the "reasonable correlation" advanced by Applicant.

The present application recites at, for example, pages 36-47 of the specification that the claimed invention can be used, *inter alia*, to identify ligands, protein binding partners, and/or modulators. Additionally, the polynucleotides of the present invention can be used to generate antibodies useful to localize proteins encoded by the polynucleotides of the present invention *in vivo* or *in vitro*. The polynucleotides can also be used to determine the expression pattern of the gene in various tissues which would enable a person of ordinary skill in the art to better understand the function and role of the gene *in vivo*. Thus, there is no question that Applicant has asserted at least one specific utility and, in fact, have provided numerous specific utilities for the polynucleotides of the present invention. Accordingly, under *Brana*, the Patent Office ***must*** accept the utility asserted by Applicant.

Additionally, the Office appears to be under the assumption that ***absolute*** certainty is required for a polynucleotide to have a specific utility. The standard applicable in this case is not, however, proof to certainty, but rather proof to reasonable probability. As the Supreme Court stated, applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner v. Manson*, 383 U.S. at 532. Although, there may be numerous inventions that may arise from the present application, this standard does not justify the Office's stance that the present invention lacks a specific utility. Thus, Applicant has complied with the specific utility requirement.

The claimed invention in *Brenner* was directed to a method whose ***only*** utility was making a class of steroids. The disclosure in Brenner failed to disclose a utility for the products of that method, which in turn led to a § 101 rejection because the products resulting from the method lacked utility. The Applicant admitted that the products produced by the method would not be patentable if they lacked utility. 148 USPQ 696. The Court stated that the method lacked utility as well, holding:

> We find absolutely no warrant for the proposition that although Congress intended that no patent be granted on a chemical compound whose sole "utility" consists of its potential role as an object of use-testing, a different set of rules was meant to apply to the process which yielded the unpatentable product.

148 USPQ 696.

In *Brenner*, the method of making the compounds, which was the only use recited, was inextricably bound up with the compounds themselves and, as a result, the requirement for utility could not be met until a use for the compounds was found. The Court emphasized that the utility of the claimed invention (i.e., the products) would require further research to identify and ascertain, and the compounds produced by the method would be the object of that research.

In contrast, GPCRs related to known GPCRs stand on a very different basis. As discussed, there are a multitude of utilities for the claimed polypeptides, including their ability to facilitate research.

Applicant furthers assert that long held pre-Brenner case law standard supports judging the utility of an invention on whether or not the public derives a benefit from the invention, regardless of how slight the benefit. *See,* for example, *In re Nelson*, 280 F.2d 172, 178-180 (C.C.P.A. 1960) (stating that "however slight the advantage which the public have received from the inventor, it offers a sufficient reason for his compensation") (citing ROBINSON ON PATENTS (1890)); *see also Lowell v. Lewis*, 1 Mason 182 (Fed. Case. No. 8568, 1817) (stating "if it be more or less useful is... of no importance to the public. If it be not extensively useful it will silently sink into contempt and disregard"). Polypeptides of all types are broadly used in the biotechnology industry, playing key roles in drug and disease discovery processes. Indeed, many such

polypeptides enable researchers to find the genes associated with physiological functions. The discovery of such functions readily benefits the public. Accordingly, such tools satisfy the pre-Brenner case law standard.

The Claimed Invention Has A Substantial Utility

The Utility Examination Guidelines also require a claimed invention to have a utility that defines a real-world use (a "substantial utility"). Applicant teaches, as described above, that the claimed invention can be used to make antibodies, identify ligands and other binding partners, such as other proteins that interact with the polypeptide (i.e., a G protein). Thus, it is clear that the claimed invention has real-world uses. All the uses described in the present application are real-world uses and, again, stand in stark contrast to the "throw away" uses (e.g., landfill component or snake food) set forth in the utility guidelines. Thus, there is no question that Applicant has asserted at least one substantial utility and, in fact, have provided numerous substantial utilities. Accordingly, Applicant has complied with the substantial utility requirement.

The Claimed Invention Has A Credible Utility

In addition to a specific and substantial utility, the Utility Examination Guidelines require that such utility be credible (a "credible utility"). That is, whether the assertion of utility is believable to a person of ordinary skill in the art based on the totality of evidence and reasoning provided. Clearly, the numerous specific and substantial utilities asserted by Applicant are credible.

Assertions of credibility are credible unless "(A) the logic underlying the assertion is seriously flawed, or (B) the facts upon which the assertion is based is inconsistent with the logic underlying the assertion." (See, Revised Interim Utility Guidelines Training Materials.) All the utilities described for the polynucleotide and polypeptide are based on sound logic. Furthermore, the utilities for the claimed polynucleotide are ***not*** inconsistent with the logic underlying the assertion that the polynucleotide are useful. Polynucleotides are useful to encode and produce

polypeptides to generate antibodies, identify ligands or protein partners, evaluate expression patterns, evaluate protein activity, etc. The Office has provided no evidence that the logic is seriously flawed or that the facts upon which these assertions are based are inconsistent with the logic underlying the assertions.

In this respect, the G protein coupled receptor family is analogous to the chemical genus that was the subject of *In re Folkers*, 145 USPQ 390 (CCPA 1965) (Compound that belongs to class of compounds, members of which are recognized as useful, is considered useful under §101.) The Patent Office does not serve the public by attempting to substitute a formulaic analysis of § 101 for the established judgment of the biopharmaceutical industry as to what is "useful." If the Patent Office is aware of any well-grounded scientific literature suggesting that GPCR's are not useful, Applicant requests that it be made of record.

Art-Recognized Utility

The Utility requirement may also be satisfied by an "Art Established Utility" which means that "a person of ordinary skill in the art would immediately appreciate why the invention is useful based on the characteristics of the invention... and the utility is specific, substantial and credible." (M.P.E.P. §2107).

Applicant points out that commercial products relating to GPCRs for which no confirmed function has been identified are commercially available. GPCRs, ORF clones of GPCRs, and antibodies that bind to GPCRs are commercially available. For example, Applicant points out that FabGennix Inc. of Shreveport, Louisiana sells an antibody directed to Retinal Anti-GP75. GPCR75 is said to be a GPCR for which a ligand has not yet been identified (*see* attached product sheet). Invitrogen sells ORF clones of GPCRs including those for which a ligand has not yet been identified (*see* attached list, especially noting Clone Ids IOH22483, IOH14039, IOH13056, IOH22637, IOH13239, and IOH13516). MD Bio of Taiwan sells GPCR peptides and antibodies against such peptides, again where no ligand has yet been identified. That at least three companies make and sell such GPCR products proves that there is a well-established utility for the

presently claimed GPCR polypeptides. Accordingly there could be no better proof of the utilities of the claimed polypeptides- such products are made by a manufacturer (who expects to sell them) for consumers (who expect to buy them). Any argument that there is no art-recognized utility for such polypeptides seems meritless.

Applicant also notes for the record that the Patent Office apparently agrees with Applicant's reasoning that GPCRs are useful in that the Office has granted and apparently continues to grant patents to G-protein coupled receptors, their encoding polynucleotides and antibodies directed to them ***in which no natural substrate or specific biological significance*** is ascribed to the GPCR. Specifically, Applicant would like to bring the following US Patents to the Office's attention:

6,518,414 Maclennan "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims an isolated polynucleotide)
6,511,826 Li et al. "Polynucleotides Encoding Human G-Protein Chemokine Receptor (CCR5) HDGNR10" (Claims an isolated polynucleotide encoding a protein identified as a "chemokine receptor" with no specific chemokine identified)
6,372,891 Soppet et al. "Human G-Protein Receptor HPRAJ70" (Claims an antibody directed to a G-protein coupled receptor)
6,361,967 Agarwal et al. "AXOR10, A G-Protein Coupled Receptor" (Claims an isolated polynucleotide)
6,348,574 Godiska et al. "Seven Transmembrane Receptors" (Claims an antibody directed to a G-protein coupled receptor)
6,114,139 Hinuma et al. "G-Protein Coupled Receptor Protein and A DNA Encoding the Receptor" (Claims an isolated polynucleotide).
6,111,076 Fukusumi et al. "Human G-Protein Coupled Receptor (HIBCD07)" (Claims isolated polypeptide)
6,107,475 Godiska et al. "Seven Transmembrane Receptors" (Claims isolated polynucleotide and methods)
6,096,868 Halsey et al. "ECR 673: A 7-Transmembrane G-Protein Coupled Receptor" (Claims isolated polypeptide)
6,090,575 Li et al. "Polynucleotides Encoding Human G-Protein Coupled Receptor GPR1" (Claims isolated polynucleotide)
6,071,722 Elshourbagy et al. "Nucleic Acids Encoding A G-Protein Coupled 7TM Receptor (AXOR-1)" (Claims an isolated polynucleotide)
6,071,719 Halsey et al. "DNA Encoding ECR 673: A 7-Transmembrane G-Protein Coupled Receptor" (Claims an isolated polynucleotide)
6,060,272 Li et al. "Human G-Protein Coupled Receptors" (Claims isolated polynucleotide)
6,048,711 Hinuma et al. "Human G-Protein Coupled Receptor Polynucleotides" (Claims isolated polynucleotide)

6,030,804 Soppet et al. "Polynucleotides Encoding G-Protein Parathyroid Hormone Receptor HLTDG74 Polypeptides" (Claims isolated polynucleotide)
6,025,154 Li et al. "Polynucleotides Encoding Human G-Protein Chemokine Receptor HDGNR10" (Claims an isolated polynucleotide encoding a protein identified as a "chemokine receptor" with no specific chemokine identified)
5,998,164 Li et al. "Polynucleotides Encoding Human G-Protein Coupled Receptor GPRZ" (Claims isolated polynucleotide)
5,994,097 Lal et al. "Polynucleotide Encoding Human G-Protein Coupled Receptor" (Claims isolated polynucleotide)
5,958,729 Soppet et al. "Human G-Protein Receptor HCEGH45" (Claims isolated polypeptide)
5,955,309 Ellis et al. "Polynucleotide Encoding G-Protein Coupled Receptor (H7TBA62)" (Claims isolated polynucleotide)
5,948,890 Soppet et al. "Human G-Protein Receptor HPRAJ70" (Claims isolated polypeptide)
5,945,307 Glucksmann et al. "Isolated Nucleic Acid Molecules Encoding A G-Protein Coupled Receptor Showing Homology to The 5HT Family of Receptors" (Claims isolated polynucleotide)
5,942,414 Li et al. Polynucleotides Encoding Human G-Protein Coupled Receptor HIBEF51" (Claims isolated polynucleotide)
5,912,335 Bergsma et al. "G-Protein Coupled Receptor HUVCT36" (Claims isolated polynucleotide)
5,874,245 Fukusumi et al. "Human G-Protein Coupled Receptors (HIBCD07)" (Claims isolated polynucleotide)
5,871,967 Shabon et al. "Cloning of A Novel G-Protein Coupled 7TM Receptor" (Claims isolated polynucleotide)
5,869,632 Soppet et al. "Human G-Protein Receptor HCEGH45" (Claims isolated polynucleotide)
5,856,443 MacLennan et al. "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims isolated polynucleotide)
5,834,587 Chan et al. "G-Protein Coupled Receptor, HLTEX11" (Claims isolated polypeptide)
5,776,729 Soppet et al. "Human G-Protein Receptor HGBER32" (Claims isolated polynucleotide)
5,763,218 Fujii et al. "Nucleic Acid Encoding Novel Human G-Protein Coupled Receptors" (Claims isolated polynucleotide)
5,756, 309 Soppet et al. "Nucleic Acid Encoding A Human G-Protein Receptor HPRAJ70 and Method of Producing the Receptor" (Claims isolated polynucleotide)
5,585,476 MacLennan "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims isolated polynucleotide)
5,759,804 Godiska et al. "Isolated Nucleic Acid Encoding Seven Transmembrane Receptors" (Claims isolated polynucleotide and methods)

Applicant asserts that these issued US Patents are evidence of an art recognized utility for G-protein coupled receptors whose natural ligand is unknown. If the Patent Office's position is that issued patents are ***not*** sufficient evidence of art recognition then Applicant respectfully requests that this position be made of record. In the alternative, if the Patent Office wishes to take the position that these issued patents are directed to non-statutory subject matter, then Applicant respectfully requests that this position be made of record.

The Office also alleges that protein belonging to the GPCR family, even if they have similar structures, can have different functions and, therefore, the invention is incomplete. However, Applicant does not determine function based on the structure of the encoded protein. Rather the prediction is based upon the sequence similarity with known polynucleotides or polypeptides encoded thereby. Although different structures can be formed by different amino acid sequences thereby allowing proteins with similar structures to have different functions, proteins that also share sequence similarity in addition to structural similarity are likely to be part of the protein family. It is well known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. Brenner et al., Proc. Natl. Acad. Sci. 95:6073-78 (1998) (See, attached reference). In the present application homology is in excess of 40% over many more than 70 amino acid residues. The probability, therefore, that the polypeptide encoded by the claimed polynucleotides is related to the reference polypeptides is, accordingly, very high.

The Office has failed to provide any references that contradict Brenner's basic rule and has failed to provide any "countervailing evidence" required by the Utility Examination Guidelines. Therefore, the Office has failed to meet its burden in providing evidence indicating that the present invention does not have a substantial, credible, and useful invention.

In view of the foregoing, Applicant respectfully requests that the rejection under 35 U.S.C. § 101 be withdrawn.

Rejections under 35 U.S.C. § 112

Claims 1-22, 25-29, and 67-72 stand rejected under 35 U.S.C. § 112, first paragraph, as allegedly failing to adequately teach how to use the instant invention. According to the Office, "Since the claimed invention is not supported by either a specific, substantial or credible utility...one skilled in the art clearly would not know how to use the claimed invention." (Office Action, page 5). Applicant respectfully disagrees.

As discussed above, the present invention ***is*** supported by a specific, substantial, and credible asserted utility as well as a well-established utility. Accordingly, Applicant respectfully requests that the rejection be withdrawn.

The Office also alleges, that "even if the specification taught how to use the nucleic acid encoding the human nGPCR-1079 polypeptide, enablement would not be commensurate in scope with claim 1 and the dependent claims 3, 5-22, 25-29, and claims 67 and 68." (Office Action, page 6). Applicant respectfully disagrees.

As presently amended, claims 1, 3, 25, and 77 recite polynucleotides that have at least 90% homology to SEQ ID NO:1 or polypeptides that having at least 95% homology to SEQ ID NO: 2.

The claims, as amended, are not excessively broad. A person of ordinary skill in the art would readily understand what is meant by "at least 90% homologous." Homology for a polypeptide and nucleic acid molecule is well understood by those of ordinary skill in the art and is described in the specification such that the present invention can be made and used by the art-skilled.

A person of ordinary skill in the art would understand that the purified and isolated polynucleotide include, for example, polynucleotides that encode for polypeptides that have mutations when compared to SEQ ID NO: 2. One of skill in the art would readily be able to make and use such polynucleotides.

Claims 1-22, 25-29, and 67-72 are also rejected under 35 U.S.C. § 112, first paragraph, as allegedly containing subject matter which was not described in the specification in such a way as to enable one skilled in the art to which it pertains, or with

which it is most nearly connected, to make and/or use the invention. The Office alleges that the polypeptide encoded by the nucleic acid molecule is not a complete sequence of a GPCR and therefore, it "is unlikely that the present nGCPR-1079 of SEQ ID NO: 2, which has merely 1/3 of the minimum length of a GPCR, is a functional GPCR even though it may be a portion of a GPCR, and undue experimentation is required prior to using the present invention for any purpose as claimed." (Office Action, page 8) Applicant respectfully disagrees.

Although the disclosed sequences may not be full-length GPCRs, there is no indication that the present sequences possess no function or that the sequences cannot be used for other purposes even if they do not retain GPCR activity. Notably, the Office has failed to provide any evidence whatsoever that a polypeptide encoded by the claimed polynucleotides do not retain GPCR activity. One of ordinary skill in the art can readily determine if the polypeptide encoded by the polynucleotide of the present invention has activity. Experiments performed to determine activity are routine and well known by one of skill in the art. The Office is respectfully reminded that the relevant issue is not the amount of experimentation, but rather whether any experiments that may be performed would be undue to one of skill in the art. Enzymatic assays are routine to those of skill in the art. Assays to measure GPCR function are also well known in the art and are also described in the present application. Therefore, one of skill in the art would know how to make and/or use the present invention.

However, even if the encoded polypeptide did not possess GPCR activity, one of skill in the art can still use the polypeptide to raise antibodies, to identify binding partners through various assays (*i.e.* yeast two-hybrid), and the like. These experiments are routine to one of ordinary skill in the art and do not impose an undue burden.

In view of the foregoing, Applicant respectfully requests that the rejection of claims under 35 U.S.C. § 112, first paragraph be withdrawn.

Claims 1-4, 8, 22, 27, 67, 69, 71, and 72 were also rejected under 35 U.S.C. § 112, first paragraph, as allegedly containing subject matter which was not described in

the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. Applicant respectfully disagrees

According to the Office, "only the isolated nucleic acid of SEQ ID NO:1 or encoding the amino acid sequence of SEQ ID NO:2, but not the full breadth of the claims meets the written description provision of 35 U.S.C. § 112, first paragraph.

Preliminarily, Applicant thanks the Office for its acknowledgement that written description support exists in the specification for nucleic acids encoding SEQ ID NO:2.

As discussed above, the claims have been amended to recite a specific level of homology. Applicant asserts that a skilled artisan can readily envision the structure of the claimed polypeptides and nucleic acid molecules based on the present application. One of ordinary skill in the art understands that the polynucleotides or the polypeptides of the present invention will have at least 90% homology to either SEQ ID NO: 1 or SEQ ID NO: 2. This is more than "a mere statement that is part of the invention". Rather, the recited percent homology is a defining structural characteristic of the present invention. The present invention encompasses only the nucleic acid molecules or polypeptides with at least 90% homology to SEQ ID NO: 1 or SEQ ID NO: 2.

New claim 80 has been added that recites a nucleic acid molecule that encodes for a polypeptide that is at least 99% homologous to SEQ ID NO:2. Applicant respectfully asserts that the skilled artisan can readily envision the detailed chemical structure of the polypeptides encompassed by new claim 80.

The subject matter encompassed by the pending claims is described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention.

Therefore Applicant respectfully requests that the rejection of claims under 35 U.S.C. § 112, first paragraph be withdrawn.

Claims 1-22, 25-29 and 72 stand rejected under 35 U.S.C. § 112, second paragraph, as allegedly indefinite for failing to particularly point out and distinctly claim

the subject matter which applicant regards as the invention. Applicant respectfully disagrees.

The Office alleges that claim 1 is indefinite because it is not clear what "homologous." According to the office, the "claim does not specific the percentage of the sequence identity or any other objective measurement." (Office Action, page 9). Applicant has amended claim 1 to recite "at least 90% homologous", rendering this rejection moot.

The Office alleges that claim 10 is indefinite for the recitation of "said vector is a viral particle". Applicant has amended claim 10 to recite "said vector is a viral vector" rendering this rejection moot.

Claim 22 stands rejected as allegedly indefinite. Applicant has canceled claim 22 without prejudice, rendering this rejection moot.

Claims 25 and 26 stand rejected as allegedly indefinite for the recitation of "an acceptable carrier or diluent" because it is allegedly unclear what is "acceptable." Applicant respectfully disagrees, but in order to further prosecution, has amended claims 25 and 26 to recite a "pharmaceutically acceptable carrier or diluent." The term "pharmaceutically acceptable" is described in the present specification (see, for example, pages 32-33) and is also well known to those of skill in the art.

Claim 27 stands rejected as allegedly indefinite for using the inclusive language "and" in "a polypeptide that comprises a sequence of SEQ ID NO 2 *and* homologs thereof." Applicant has amended claim 27 removing the phrase "and homologs thereof" rendering this rejection moot.

In view of the foregoing, Applicant respectfully requests that the rejection under 35 U.S.C. § 112, second paragraph be withdrawn.

Rejections under 35 U.S.C. § 102 and § 103

The Office rejected claims 1-22, 25-29, and 67-71 under 35 U.S.C. § 102 and/or § 103 in view of its erroneous assertion that the effective filing date for the instantly

claimed invention is August 15, 2001, which is the actual filing date of the instant application.

As discussed above, the effective filing date of the present application is that of its priority application, filed August 15, 2000. The Office alleges that the present application is not entitled to its priority date because the prior application did not satisfy the requirements under 35 U.S.C. § 101 and 112, first paragraph. However, as discussed above, the prior application does satisfy the requirements under 35 U.S.C. § 101 and § 112, first paragraph, for the reasons set forth above, and therefore is entitled to the priority date of August 15, 2000.

Claims 1-22, 25-29, and 67-71 stand rejected under 35 U.S.C. § 102(e) as allegedly anticipated by Paszty et al (US 20002/0123618. The effective date of Paszty is August 10, 2001, which is after the effective date of the present application (August 15, 2000). Therefore, the Paszty reference does not qualify as prior art against the present application.

In view of the foregoing, Applicant requests that the rejection under 35 U.S.C. § 102(e) be withdrawn.

Claims 1-9, 13, 16, 20-22, 25, 26, and 69-71 stand rejected under 35 U.S.C. § 102(a) as allegedly anticipated by Chen et al. (WO 01/36471). The effective date of Chen is May 25, 2001, which is after the effective date of the present application (August 15, 2000). Therefore, the Chen reference does not qualify as prior art against the present application.

In view of the foregoing, Applicant requests that the rejection under 35 U.S.C. § 102(e) be withdrawn.

Claims 10-12, 14, 15, 17-19, 27-29 stand rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over Chen in view of Glucksmann et al (U.S. Patent No. 5,945,307). Applicant respectfully disagrees.

As discussed above the Chen reference does not qualify as prior art for the present application. Therefore the only remaining reference is the Glucksmann reference. The Glucksmann reference discusses isolated nucleic acid molecules encoding a G-protein coupled receptor showing homology to the 5HT family of receptors. However, the Glucksmann reference fails to teach or even suggest SEQ ID NO: 1 or SEQ ID NO:2. Therefore, a person of ordinary skill in the art would not have been motivated to use SEQ ID NO: 1 or 2 and combine it with what is discussed in Glucksmann. Furthermore, even if one of skill in the art were motivated to use the Glucksmann reference, a person of ordinary skill in the art would not be in possession of the present invention because it does not teach or suggest the sequences of the present invention. Therefore, the present invention is not obvious in view of the Glucksmann reference.

In view of the foregoing, Applicant requests that the rejection under 35 U.S.C. § 103(a) be withdrawn.

Conclusion

Applicant believes the claims are in condition for allowance. An early Notice of Allowance is therefore earnestly solicited. Applicant invites the Examiner to contact the undersigned at (215) 665-6928 to clarify any unresolved issues raised by this response.

Respectfully submitted,

Daniel M. Scolnick, Ph.D.
Reg. No. 52,201

Date: November 19, 2003

COZEN O'CONNOR, P.C.
1900 Market Street
Philadelphia, PA 19103-3508
Telephone: (215) 665-2000
Facsimile: (215) 665-2013

Attachments: Marchese et al., Trends Pharmacol. Sci., 20(9):370-5, 1999
Brenner et al. Proc. Natl. Acad. Sci. 95:6073-78 (1998)
Product Sheet for Anti-GPCR-75 Antibodies
Product sheet for GPCR control peptides and antibodies (MD Bio)
Product sheet for GPCR ORF clones (Invitrogen)

Online Catalog - Invitrogen Clones

Ultimate™ ORF Browser:

Advanced Search for Ultimate™ ORF Clones

33 total records for G-Protein Coupled Receptors

Buy	Clone ID	Species	Definition	Gene Symbol
	IOH3294	Human	complement component 5 receptor 1 (C5a ligand); complement component-5 receptor-2 (C5a ligand)	C5R1
	IOH12614	Human	purinergic receptor P2Y, G-protein coupled, 11	P2RY11
	IOH22483	Human	clone MGC:33224 IMAGE:5267661, mRNA, complete cds.	RDC1
	IOH14039	Human	Similar to putative nuclear protein ORF1-FL49	ORF1-FL49
	IOH11484	Human	glycoprotein Ib (platelet), alpha polypeptide	GP1BA
	IOH1987	Human	tachykinin receptor 1 isoform short; NK-1 receptor; Tachykinin receptor 1 (substance P receptor; neurokinin-1 receptor); tachykinin 1 receptor (substance P receptor, neurokinin 1 receptor); neurokinin 1 receptor	TACR1
	IOH13056	Human	similar to POSSIBLE GUSTATORY RECEPTOR CLONE PTE01	LOC11513:
	IOH9916	Human	coagulation factor II (thrombin) receptor-like 1	F2RL1
	IOH9624	Human	vasoactive intestinal peptide receptor 2	VIPR2
	IOH10679	Human	endothelin receptor type A	EDNRA
	IOH22637	Human	Similar to parathyroid hormone receptor 1, clone MGC:34562 IMAGE:5180885, mRNA, complete cds.	PTHR1
	IOH13583	Human	Duffy blood group	FY
	IOH4585	Human	cholecystokinin B receptor	CCKBR
	IOH11033	Human	endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4; G protein-coupled receptor; LPA receptor EDG4; Lysophosphatidic acid receptor EDG4	EDG4
	IOH10866	Human	CD97 antigen isoform 2 precursor; leukocyte antigen CD97; seven-span transmembrane protein	CD97
	IOH22632	Human	formyl peptide receptor-like 1; lipoxin A4 receptor (formyl peptide receptor related)	FPRL1
	IOH22669	Human	adrenomedullin receptor	ADMR
	IOH13239	Human	super conserved receptor expressed in brain 3	SREB3

Customer Service: 1800 786 1236
Technical Support: 318 219 1123
Fax: 318 798 1849
Info@fabgennix.com
www.fabgennix.com

New Item

FabGennix Inc.
INTERNATIONAL

Novel Orphan retinal G-protein coupled Receptor (GPCR-75) selective antibodies

Anti-GPCR-75Antibodies (GPCR75-100P, GPCR75-101AP and GPCR75-112AP)

Recently a novel human G-protein coupled receptor gene has been characterized and mapped to chromosome 2p16. This gene codes for a 540 amino acid protein in retinal pigment epithelium (RPE) and cells surrounding retinal arterioles. In contrast, the Northern blot data obtained from mouse sections suggest the expression of transcripts in photoreceptor inner segments and I outer plexiform layer. The transcripts of the GPCR-75 gene (7kb) are also found in abundance in brain sections. So far, no mutations in GPCR-75 protein were identified in patients suffering from Doyne's honeycomb retinal dystrophy (DHRD), an inherited retinal degeneration disease that maps to chromosome 2p16 (1).

The GPCR-75 protein is approximately 78 kDa (540 amino acids) protein that is primarily expressed in human retinal pigment epithelium (RPEs). The GPCR-75 sequence analyses suggest the presence of 7 trans-membrane domains, a characteristic feature of GPCR. The protein has putative N-glycosylation sites near the extra cellular N-terminal end of the proteins. The protein has a large 3 intra cellular loop which might be the site for interaction of G-proteins. The short carboxy terminal is intracellular and has putative post-translational modification lipid modification sites.

The Anti-GPCR-75-selective antibodies were generated against conserved sequences near N- and C-termini of the protein that are unique to GPCR-75 protein. The polyclonal antibody strongly labels a 78 kDa protein in RPE cell extracts. Anti-GPCR-75-selective antibody is also available in affinity-purified form for confocal, Western blotting and immunocytochemical analyses. *FabGennix Int. Inc.* will also conjugate antibodies with fluorescent probes upon request at extra charge. *FabGennix Int. Inc.* will also provides antibodies against proteins that are involved in retinal degenerative diseases such as various Anti-PDE antibodies, Anti-MERTK, Anti-Phospho-MERTK, EGF-containing fibulin like intracellular protein (EFEMP1), Anti-Myocilin (TIGR), Anti-Bestrophin, Anti-ELVOL4 and a Usher syndrome specific Anti-USH2a antibodies etc. FabGennix Int. Inc employs cyclic peptide methodology for generating antibodies, which results in higher titer and specificity (2). FabGennix Int. Inc., will also provide Western blot positive controls for most of these antibodies in ready-to-use buffer for easy identification of respective proteins. Limited quantities of antigens are also available. Please enquire for their availability before ordering.

Catalog #	Host Species	Nature	Cross reactivity	Quantity	volume	Price
GPCR75-100P	Rabbit	Polyclonal antisera	R, M, H	100 ml	100 ul	$ 195.00
GPCR75-101AP	Rabbit	Affinity purified IgG	R, M, H	100 ug	150 ul	$ 225.00
GPCR75-112AP	Rabbit	Affinity purified IgG	R, M, H	100 ug	150 ul	$ 225.00
PC-GPCR75	N/A	WB positive control	Rat	For 5 App	60 ul	$ 75.00
P-GPCR75	N/A	Antigenic peptides	n/a	250 ug	inquire	$ 65.00

R = rat; M = mouse; H = human; C = chicken; monk = monkey ; * not all variants are labeled equally

Immunogen: Synthetic cyclic peptide (GPCR75-101AP = PNATSLHVPHSQEGNSTS-amide; GPCR75-112AP = STSLQEGLQDLIHTATLVTC-amide).

Concentration: GPCR75-101AP; GPCR-112AP IgG concentration 0.75-1.25 mg/ml in 50% antibody stabilization buffer.

Applications: Antibody GPCR75-100/GPCR75-101AP are ideal for WB, IMM and IHC assays. The dilutions for this antibody is for reference only, investigators are expected to determine the optimal conditions for specific assay in his/her laboratory. Dilutions: WB > 1:500; Immunoprecipitation & i.p pull-down assays> 1:250

Reactivity: This antibody detects a single 78 kDa Orphan GPCR75 protein in human RPE cell extracts.

Protocols: Standard protocol for various applications (WB; IMM and IHC) of this antibody is provided with the product specification sheet, however, FabGennix Int. Inc. strongly recommends investigators to optimize conditions for use of this antibody in their laboratories.

Form/Storage: The antiserum is supplied in antibody stabilization buffer with 0.02% sodium azide or thimerosal/merthiolate as preservative. The affinity-purified antibodies are purified on antigen-sepharose affinity column and supplied as 1-1.25 mg/ml IgG in antibody stabilization buffer containing preservatives with low viscosity and cryogenic properties. For long-term storage of antibodies, store at -20°C. Now these antibodies can be stored at –20°C and used immediately with out thawing. FabGennix Inc. does not recommend storage of very dilute antibody solutions unless they are prepared in specially formulated multi use antibody dilution buffer (Cat # DiluOBuffer). Working solutions of antibodies in DiluOBuffer should be filtered through 0.45μ filter after every use for long-term storage.

78 kDa Orphan Receptor-75 in human RPE cells. Antibody GPCR-100P (1:400)

References:

1. Tarttelin E. E., Krischner L. S., Bellingham J., Baffi. J. Taymanas S. E., Gregor E. K., Csaky K., Stratakis C. A., Gregory-Evans C. Y. Biochem. Biophys. Res. Commun. 260, 174-180, 1999.
2. Farooqui, S. M., Brock. W. J., A. Hamdi., Prasad. C. (1991) J. Neurochem. 57, 1363-1369.

* For users who may require large amounts of GPCR75-100P or GPCR75-101AP, please enquire about bulk material discounts.
This Product is for Research Use Only and is NOT intended for use in humans or clinical diagnosis.

061901-0020SF1001Z-rev10.00

FabGennix Inc.
INTERNATIONL
2940 Youree Drive, Suite E, Shreveport, LA 71104

生工有限公司

Rat Taste Receptor 2 (TR2) Antibodies

Rat Taste Receptor 2 (TR2) Antibodies

Cat. # TR21-P, Rat TR2 Control Peptide # 1, SIZE: 100 ug/100 ul
FORM: Œ Soln Œ Lyophilized Lot # 3113P

Cat. # TR21-S, Rabbit Anti-rat TR2 antiserum # 1, SIZE: 100 ul neat antiserum
FORM: Œ Soln Œ Lyophilized. Lot # 38889S

Cat. # TR21-A, Rabbit Anti-rat TR2 Ab # 1 (affinity pure) SIZE: 100 ug
FORM: Œ Soln Œ Lyophilized. Lot # 38889A

Higher vertebrates are believed to possess at least five basic tastes: Sweet, bitter, sour, salty, and unami (the taste of monosodium glutamate). Taste receptor cells that may selectively reside in various parts of the tongue and respond to different tastants and perceive these taste modalities. Circumvallate papillae, found at the very back of the tongue, are particularly sensitive to biter substances. Foliate papillae, found at the posterior lateral edge of the tongue, are sensitive to sour and bitter. Fungiform papillae at the front of the tongue specialize in sweet taste.

Recently, two novel taste receptors, TR1 and TR2, have been cloned with distinct topographical distribution in taste receptor cells and taste buds. TRs are members of a new group of 7 TM domain containing GPCR distantly related to other chemosensory receptors (Ca+-sensing receptor (CaSR, a family of putative hormone receptor (V2R), and metabotropic glutamate receptors). TR1 is expressed in all fungiform taste buds, whereas TR2 localized to the circumvallate taste buds. Both receptors do not co-localize with gustducin.

Source of Antigen and Antibodies

TR1 (rat 840 aa) and TR2 (rat 843 aa) share ~40% homology with each other, and ~30% with CaSR, and 22-30% with V2R pheromone receptors and mGLURs. Rat TR are 7 TM domain containing protein with an extra long N-terminal, extracellular domain (1). A 19 AA Peptide (designated TR21-P; control peptide) sequence near the C-terminus of rat TR2(1) was selected for antibody production. The peptide was coupled to KLH, and antibodies generated in rabbits. Antibody has been affinity purified using control peptide-Sepharose.

Form & Storage

Control peptide Solution is provided in PBS, pH 7.4 at 1 mg/ml (100 ug/100 ul). Antiserum is supplied as neat serum (100 ul soln or lyophilized). Affinity pure antibodies were purified over the peptide-Sepharose column and supplied as 1 mg/ml soln in PBS, pH 7.4 and 0.1% BSA as stabilizer (100 ul in solution or Lyophilized).
The peptides and antibodies also contain 0.1% sodium azide as preservative. Lyophilized products should be reconstituted in 100 ul water and gently mixed for 15 min at room temp. All peptide/antibody

received in solution or

reconstituted from lyophilized vials should be stored frozen at -20oC or below in suitable aliquots. It is not recommended to store diluted solutions. Avoid repeated freeze and thaw.

Recommended Usage

Western Blotting (1:1K-5K for neat serum and 1-10 ug/ml for affinity pure antibody using ECL technique).

ELISA: Control peptide can be used to coat ELISA plates at 1 ug/ml and detected with antibodies (1:10-50K for neat serum and 0.5-1 ug/ml for affinity pure).

Histochemistry & Immunofluorescence: We recommend the use of affinity purified antibody at 1-20 ug/ml in paraformaldehyde fixed sections of tissues (1).

Specificity & Cross-reactivity

The 19 AA rat TR21-P control peptide is specific for rat TR2. It has no significant sequence homology with TR1 or gustducin or pheromone receptors. Antibody cross-reactivity in various species has not been studied. The TR21-P control peptide is available to confirm specificity of antibodies.

References:

1. Hoon MA et al (1999) Cell 96, 541-555; Lindemann B (1999) Nature Med. 5, 381-382

"Neat Antisera" are the unpurified antiserum and it is suitable for ELISA and Western.
"Affinity pure" antibodies have been over the antigen-affinity column and recommended for immunohistochemical applications.
"Control peptides" can not be used for Western as they are very short peptides. They are intended for ELISA or antibody competition studies.

List of Related Products

回頁首 回首頁

生工有限公司
MDBio, Inc.

全省免付費服務專線：0800-072-222 | 電話：(02)27474876 | 傳真：(02)28238024

網站：www.mdbio.c m.tw | 電子郵件：mdbio@mdbio.com.tw

中華民國92年06月09日更新

64 Na, S. et al. (1996) J. Biol. Chem. 271, 11209–11213
65 Danley, D. E., Chuang, T-H. and Bokoch, G. M. (1996) J. Immunol. 157, 500–503
66 Mills, J. C., Stone, N. L., Erhardt, J. and Pittman, R. N. (1998) J. Cell Biol. 140, 627–636
67 Subauste, M. C. et al. J. Biol. Chem. (in press)
68 Billadeau, D. D. et al. (1998) J. Exp. Med. 188, 549–559
69 Rudel, T. and Bokoch, G. M. (1997) Science 276, 1571–1574
70 Cardone, M. H., Salvesen, G. S., Widmann, C., Johnson, G. and Frisch, S. M. (1997) Cell 90, 315–323
71 Chuang, T-H., Hahn, K. M., Lee, J-D., Danley, D. E. and Bokoch, G. M. (1997) Mol. Biol. Cell 8, 1687–1698
72 Lores, P., Morin, L., Luna, R. and Gacon, G. (1997) Oncogene 15, 601–605
73 Ward, C. et al. (1999) J. Biol. Chem. 274, 4309–4318
74 Sulciner, D. J. et al. (1996) Mol. Cell. Biol. 16, 7115–7121
75 Perona, R. et al. (1997) Genes Dev. 11, 463–475
76 Hirshberg, M., Stockley, R. W., Dodson, G. and Webb, M. R. (1997) Nat. Struct. Biol. 4, 147–152
77 Krengel, U. et al. (1990) Cell 62, 539–548
78 Ihara, K. et al. (1998) J. Biol. Chem. 273, 9656–9666
79 Abdul-Manan, N. et al. (1999) Nature 399, 379–383
80 Mott, H. R. et al. (1999) Nature 399, 384–388
81 Wu, W. J., Leonard, D. A., Cerione, R. A. and Manor, D. (1997) J. Biol. Chem. 272, 26153–26158
82 Fujisawa, K. et al. (1998) J. Biol. Chem. 273, 18943–18949
83 Scheffzek, K. et al. (1997) Science 277, 333–338
84 Scheffzek, K., Ahmadian, M. R. and Wittinghofer, A. (1998) Tr Biochem. Sci. 23, 257–262
85 Sprang, S. R. and Coleman, D. E. (1998) Cell 95, 155–158
86 Boriack-Sjodin, P. A., Margarit, S. M., Bar-Sagi, D. and Kuriya (1998) Nature 394, 337–343
87 Peyroche, A. et al. (1999) Mol. Cell 3, 275–285
88 Chardin, P. and McCormick, F. (1999) Cell 97, 153–155
89 Gibbs, J., Oliff, A. and Kohl, N. E. (1994) Cell 77, 175–178
90 Uehata, M. et al. (1997) Nature 389, 990–994
91 Kumar, C. C. et al. (1995) Cancer Res. 55, 5106–5117
92 Walsh, A. B., Dhanasekaran, M., Bar-Sagi, D. and Kumar, C. C. (1 Oncogene 15, 2553–2560
93 Morozov, I., Lotan, O., Joseph, G., Gorzalczany, Y. and Pick, E. (1 J. Biol. Chem. 273, 15435–15444
94 Kreck, M. L., Uhlinger, D. J., Tyagi, S. R., Inge, K. L. and Lamb J. D. (1994) J. Biol. Chem. 269, 4161–4168
95 Heyworth, P. G., Knaus, U. G., Settleman, J., Curnutte, J. T. Bokoch, G. M. (1993) Mol. Biol. Cell 4, 1217–1223
96 Reif, K., Nobes, C. D., Thomas, G., Hall, A. and Cantrell, D. A. (1 Curr. Biol. 6, 1445–1455
97 Zhou, K. et al. (1998) J. Biol. Chem. 273, 16782–16786
98 Manser, E. et al. (1998) Mol. Cell 1, 183–192

Acknowledgements
The authors were supported by grants from the National Institutes of Health (GMB), The National Arthritis Foundation, and the Association pour la Recherche sur le Cancer (VB). We wish to thank Dr Celine Der Mardirossian for thoughtful suggestions about the text and Antonette Lestelle for secretarial assistance. Suggestions offered by the reviewers of this manuscript, which greatly improved the organization and content, are gratefully acknowledged.

Novel GPCRs and their endogenous ligands: expanding the boundaries of physiology and pharmacology

Adriano Marchese, Susan R. George, Lee F. Kolakowski Jr, Kevin R. Lynch and Brian F. O'Dowd

A. Marchese, Graduate Student, Dept of Pharmacology, Email: a.marchese@utoronto.ca
S. R. George*, Professor, Depts of Pharmacology and Medicine, Email: s.george@utoronto.ca
and **B. F. O'Dowd*,** Associate Professor, Dept of Pharmacology, University of Toronto, Medical Sciences Building, Toronto, ON, Canada M5S 1A8. Email: brian.odowd@utoronto.ca
*Also at Center for Addiction and Mental Health, 33 Russell Street, Toronto, ON, Canada M5S 2S1.
L. F. Kolakowski Jr, Associate Professor, Dept of Pharmacology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78284-7763, USA. Email: kolakowski@uthscsa.edu
and **K. R. Lynch,** Professor, Dept of Pharmacology, University of Virginia Health Sciences Center, 1300 Jefferson Park Ave, Charlottesville, VA 22908, USA. Email: krl2r@virginia.edu

Nearly all molecules known to signal cells via G proteins have been assigned a cloned G-protein-coupled-receptor (GPCR) gene. This has been the result of a decade-long genetic search that has also identified some receptors for which ligands are unknown; these receptors are described as orphans (oGPCRs). More than 80 of these novel receptor systems have been identified and the emphasis has shifted to searching for novel signalling molecules. Thus, multiple neurotransmitter systems have eluded pharmacological detection by conventional means and the tremendous physiological implications and potential for these novel systems as targets for drug discovery remains unexploited. The discovery of all the GPCR genes in the genome and the identification of the unsolved receptor–transmitter systems, by determining the endogenous ligands, represents one of the most important tasks in modern pharmacology.

The G-protein-coupled receptors (GPCRs) are transdu of extracellular messages and they allow tissues to resp to a wide array of signalling molecules. Most of the dogenous ligands are small and the binding of these gands to their receptor(s) can be mimicked (or block by synthetic analogues. Together with the knowledge numerous GPCRs are targets of important drugs in today, GPCR identification is a task of prime importa In the 14 years since the first cloning of genes for GPC most of the molecules known to signal cells via the het trimeric G-protein–effector systems have been assig a cloned GPCR gene. However, the vigorous search novel GPCR genes has far outpaced the identificatio novel endogenous ligands. A group of genes has been id tified whose products are, using the criterion of seque similarity, members of the GPCR family but for which ligands are not known, and these are commonly kn as orphans (oGPCR).

The GPCR gene family is the largest known rece family (see Box 1) and shares a common secondary st ture that consists of seven transmembrane domains. ting aside the odorant receptors (encoded by hund of genes), nearly 300 mammalian GPCR genes have recognized[1]. On the basis of structure, the GPCRs ca separated into three subfamilies. The inclusion of a re tor in a subfamily requires the presence of an overall centage amino acid identity and not any discrete m Most GPCRs, including the odorant receptors, are grou in Family A. Several additional GPCRs, which hav their ligands peptides such as secretin, vasoactive int nal peptide and calcitonin, make up Family B. Fami comprises the metabotropic glutamate receptors, Ca^{2+}-sensing receptor, pheromone receptors, the GA receptors and the taste receptors. Within each fa GPCRs are grouped by sequence similarity and li specificity; approximately one third of Family A men

Box 1. How big is the GPCR family?

The size of the GPCR family surprised even the most optimistic pharmacologist as many subfamilies proved to be larger than had been predicted by classical pharmacological techniques. Furthermore, some ligands that were not widely considered to signal via receptors (e.g. nucleotides) are recognized now to have numerous receptor subtypes. The discovery of these multiple subtypes, new ligands and the rapid accumulation of novel GPCR sequences have led to the expectation that many more mammalian GPCRs await discovery. Thus, an obvious question to ask is: how many GPCR genes are there in the human genome? Although simply waiting a few years should answer this question directly, there are practical implications in making an educated guess now. For example, is the receptor for a candidate ligand likely to be visible now among the existing oGPCR DNAs? And, is further searching for oGPCR DNAs a worthwhile endeavour?

The recent completion of the nematode (*Caenorhabditis elegans*) translated genome provides an interesting comparison to mammalian GPCRs. In contrast to the single cell yeast (with its two GPCR genes), multicellularity obviously demands cell-to-cell communication and the added complexity imposes a requirement for a much larger repertoire of GPCRs. According to th analysis reported by Bargmann[1], 5% of the 19 100 nematode genes encode GPCRs. Their distribution among GPCR families is reminiscent of the mammalian GPCR genes, some 700–1000 chemoattractant (odorant) genes (including numerous pseudogenes), approximately 150 Family A genes and four-to-five each Family B and C genes. By analogy, this suggests that the number of mammalian GPCRs could total 5000 (5% of mammalian genes estimated to be 80 000–100 000). Unfortunately, the *C. elegans* genome provides no direct clues for oGPCR identification as the closest nematode GPCR is <35% identical to any mammalian GPCR and there are no obvious homologues to mammalian pre-pro-neuropeptide genes. In contrast, the accumulation of nucleotide sequence information from another surrogate organism, the zebrafish (*Danio rerio*), should be more informative because the conceptualized GPCR amino acid sequences are often ~70% identical to orthologous mammalian GPCRs.

Reference

1 Bargmann, C. (1998) *Science* 282, 2028–2033

are oGPCRs and this review will focus on these receptors. Thus, in a decade, the list of signalling molecules for which the GPCR genes had not been cloned has been supplanted by a list of ~80 oGPCRs awaiting a ligand (see Table 1). The characterization of these GPCRs has already enabled the discovery of several new endogenous ligands; this will be discussed later.

Novel GPCR gene discovery

Very few GPCRs have been purified, thus the pace of GPCR gene discovery has been fuelled by a series of highly successful cloning techniques. The identification (using amino acid sequence determination and expression cloning) of a few sequences encoding Family A GPCRs demonstrated that these were related genes[1]. Cloning by low stringency hybridization to cDNA/genomic DNA libraries yielded a stream of novel GPCR DNAs. The pace of discovery quickened with the use of the polymerase chain reaction (PCR). The database of expressed sequence tagged cDNAs (ESTs) has provided material for a further expansion of Family A, as has the high-throughput sequencing of 100–200 kb pair segments of human DNA.

Novel GPCR identification

Many oGPCRs are found to be similar to known GPCRs. Where the identity reaches the threshold of ~45%, it is likely that the receptors will share a common ligand, i.e. that the oGPCR will be a pharmacological subtype of the known GPCR. This rule is not without exception. Take, for example the orphanin FQ/nociceptin receptor; this has ~65% amino acid identity to opioid receptors, but does not have high affinity for opioid peptides[2,3]. Many GPCR subtypes have <40% amino acid identity, in which case sequence comparison might not be profitable. Moreover, because the ligand-binding pocket has not yet been described fully for any receptor, it is not feasible to predict ligand identity. However, dendritic tree building shows that receptors that respond to the same, or similar, agonists often cluster. For example, most members of the prostanoid receptor subfamily share <30% amino acid identity, yet these eight receptors are more like one another than any other GPCR. A similar situation exists among the nucleotide receptors, chemokine receptors and other cationic amine receptors. In the way that many known GPCRs fall into subfamilies, many oGPCRs cluster together, sometimes with members having >50% amino acid identity, which suggests that the problem of the ~80 oGPCRs might be solved by a mere 30 or 40 ligands. For example, the recent identification of Edg-1 as a sphingosine 1-phosphate receptor[4–6] leads directly to the prediction that Edg-3 and Edg-5 (both >50% identical to Edg-1) have the same ligand. More distant members of the Edg cluster, Edg-2 and Edg-4 are known to be receptors for the structurally related ligand, lysophosphatidic acid[7–9].

When homology does not inform, i.e. the nearest known GPCR has <35% amino acid identity to the orphan, ligand identification is challenging. There are no signature amino acids that predict either the nature of the ligand or the identity of the interacting Gα subunit type(s). In those cases where the ligand is a molecule with an established pharmacology, tissue distribution has allowed inference of ligand identity. Thus, an important clue to identifying the oGPCR RDC-8 as encoding the adenosine A_{2A} receptor was the concordance of *in situ* hybridization and ligand ([^{3}H]CGS21680) autoradiography signals in rat brain sections[10]. Similarly, the occurrence of both cannabinoid binding sites and SKR6 receptor mRNA accumulation in NG108 cells led to the identification of the cannabinoid CB_1 receptor[11].

Table 1. Amino acid sequence identity of some orphan G-protein-coupled receptors

Homology	Name	Species	% Amino acid identity	Accession no.
Opioid and somatostatin receptor-like	GPR7	Human	62% GPR8, 40% sst_5	U22491
	GPR8	Human	62% GPR7, 45% sst_5	U22492
	GPR24	Human	33% sst_2, 32% sst_5	U71092
	GPR14	Rat	29% μ-opioid, 28% sst_4	U32673
	GPR54	Rat	37% gal2, 35% GAL1	AF115516
Chemokine receptor-like	GPR2	Human	41% CXCR3, 40% CCR7	U13667
	CKRX	Human	53% EO1, 43% CCR1	AF014958
	EO1	Mouse	53% CKRX, 36% CCR1	AF030185
	MIP-1αRL1	Mouse	62% CCR1, 50% CCR3	U28405
	GPR28	Human	43% CCR7, 38% CCR6	U45982
	STRL33	Human	37% CCR7, 37% CCR6	U73529
	PPR1	Bovine	39% CCR7, 37% GPR28	S63848
	g10d	Rat	33% RDC1, 30% CCR9	L09249
	RDC1	Human	33% g10d, 30% CXCR2	X14048
	TM7SF1	Human	22% GPR5, 14% CCR6	AF027826
	CLR1	Chicken	51% BLR1, 36% CXCR1	AF029369
Chemoattractant receptor-like	Dez	Human	37% GPR1, 35% FPR2	U79527
	FPRL2	Human	72% FPR2, 56% FPR1	M76673
	FPR2	Human	72% FPRL2, 69% FPR1	M76672
	GPR1	Human	37% Dez, 34% FPR2	U13666
	GPR30	Human	32% FPRL2, 32% FPR2	AF027956
	GPR32	Human	39% FPR1, 35% FPRL2	AF045764
	GPR33	Mouse	36% GPR32, 36% Dez	AF045766
	GPR44	Human	37% Dez, 36% FPRL2	AF118265
	mas oncogene	Human	34% MRG, 26% C5aR	M13150
	MRG	Human	34% *mas* oncogene, 34% C5aR	S78653
	RTA	Rat	32% *mas* oncogene, 33% MRG	M32098
	GPR53p	Human	35% MRG, 28% *mas* oncogene	AF096785
Angiotensin receptor-like	GPR15	Human	34% GPR25, 31% APJ	U34806
	GPR25	Human	34% GPR15, 32% APJ	U91939
Cannabinoid receptor-like	GPR3	Human	59% GPR6, 57% GPR12	U13668
	GPR6	Human	59% GPR3, 56% GPR12	L36150
	GPR12	Rat	57% GPR3, 56% GPR6	U18548
	EDG-6	Human	46% EDG-3, 44% EDG-1	AJ000479
GPR4 receptor-like	OGR1	Human	48% GPR4, 35% TDAG8	U48405
	GPR4	Human	48% GPR12A, 36% TDAG8	L36148
	TDAG8	Human	36% GPR4, 35% GPR12A	U95218
	G2A	Mouse	34% GPR4, 31% OGR1	AF083442
Neuropeptide Y receptor-like	GIR	Mouse	35% GPR10, 30% NK_4	M80481
	GPR19	Human	27% GAL1, 26% NPY Y_2	U64871
	GPR22	Human	26% NPY Y_6, 24% CCK_1	U66581
Amine receptor-like	PNR	Human	33% 5-HT_4, 33% 5-HT_7	AF021818
	GPR26	Human	28% 5-HT_{5B}, 23% 5-HT_{5A}	
	GPR27	Mouse	29% D4, 25% 5-HT_6	AF027955
	AGR9	Rat	24% H_2, 24% NK_2	S73608
	GPR21	Human	27% β_3AR, 24% β_2AR	U66580
	PSP24	Human	26% 5-HT_4, 23% β_1AR	U92642
	GPR45	Human	70% PSP24, 21% NK_2	AF118266
	A-2	Human	21% 5-HT_{1F}, 19% 5-HT_{1E}	U47928
	GPR52	Human	71% GPR21, 27% H_2	AF096784
	RE2	Human	25% α_{1A}AR, 25% α_{1C}AR	AF091890
	GPR57	Human	59% GPR58, 37% PNR	N/A
	GPR58	Human	59% GPR57, 42% PNR	N/A
	GPR61	Human	27% LZY2, 30% 5-HT_6	N/A
	GPR62	Human	27% LZY, 28% 5-HT_6	N/A
P2 receptor-like	GPR23	Human	53% RBintron, 33% '$P2Y_{10}$'	U66578
	RBintron	Human	53% GPR23, 38% $P2Y_4$	L11910
	GPR35	Human	32% GPR23, 30% HM74	AF027957
	'$P2Y_{10}$'	Human	34% RBintron, 33% GPR23	AF000545
	GPR17	Human	35% $P2Y_2$, 34% $P2Y_4$	U33447
	GPR18	Human	30% RBintron, 29% GPR17	L42324
	HM74	Human	36% GPR31, 29% $P2Y_1$	D10923
	GPR31	Human	36% HM74, 29% $P2Y_1$	U65402

Table 1. (*cont.*)

Homology	Name	Species	% Amino acid identity	Accession no.
P2 receptor-like (*cont.*)	RSC338	Human	33% H963, 28% tp2y	D13626
	EBI 2	Human	33% RBintron, 30% CCR1	L08177
	H963	Human	33% RSC338, 28% PAFR	AF002986
	GPR41	Human	98% GPR42, 41% GPR43	AF024688
	GPR42	Human	98% GPR41, 28% GPR23	AF024689
	GPR40	Human	31% GPR43, 26% CXCR1	AF024687
	GPR43	Human	41% GPR41, 31% GPR40	AF024690
	GPR20	Human	31% $P2Y_4$, 26% GPR23	U66579
	GPR34	Human	31% RSC338, 29% RBintron	AF118670
	GPR55	Human	29% $P2Y_5$, 30% GPR23	AF096786
Neurotensin receptor-like	GHS-R	Human	35% NTS1, 33% nts2	U60179
	GPR39	Human	32% NTS1, 25% nts2	AF034633
	HSOGPCR2	Human	38% GPR38, 34% GHS-R	AF044601
Melatonin receptor-like	H9	Human	48% ML_{1A}, 45% ML_{1B}	U52219
Endothelin receptor-like	GPR37	Human	68% ET_BR-LP-2, 27% ET_B	U87460
	ETBR-LP-2	Human	68% GPR37, 27% ET_B	Y16280
Glycoprotein hormone receptor-like	LGR5	Human	26% FSH-R, 25% LH-R	AF062006
Opsin receptor-like	Encephalopsin	Human	32% Peropsin, 31% Rhodopsin	AF140242
	RGR	Human	27% Peropsin, 26% Rhodopsin	U15790

Please refer to the *TiPS Receptor and Ion Channel Nomenclature Supplement* and to individual GenBank accession numbers for further information.

Endogenous ligand identification

In the same way that EST database searching has yielded GPCR DNAs, it has also yielded DNAs encoding peptide sequences related to known peptides. Several novel chemokines have been discovered using this approach and these have proven to be the ligands for several chemokine receptors. For example, a CC chemokine termed ELC (EBI-ligand chemokine) was identified from the EST database and found to be the endogenous ligand for the orphan receptor EBI1, which has since been renamed CCR7 (Ref. 12). Similarly, the CC chemokine liver and activation-regulated chemokine (LARC) was identified from the EST database[13] and subsequently shown to be the ligand for the orphan STRL22 receptor; this was renamed CCR6 (Refs 14–16). Another EST encoding a CXC chemokine was isolated, BCA1 (Ref. 17), and later identified as a ligand for the oGPCR BLR1, which has since been renamed CXCR5 (Ref. 18). A fourth, novel class of chemokines called δ-chemokines, or CX_3C chemokines, was discovered by automated high-throughput single-pass sequencing and analysis of a cDNA library constructed from murine choroid plexus[19]. The sequence of one of the cDNA clones exhibited similarity to murine monocyte chemoattractant protein-1 (MCP-1), an α-chemokine. Also, another group independently searched the EST database with known chemokine sequences and identified the same chemokine, which they have termed fractalkine[20]. This ligand was matched to the orphan receptor V28 (renamed CX3CR1)[21]. The ligand for the novel receptor encoded by GPR5 (Ref. 22) has been identified as the single C motif-1 peptide[23] and the receptor renamed as XC chemokine receptor I. The ongoing search for the discovery of novel chemokines will most certainly reveal novel candidates to test with the existing chemokine-like orphan receptors and any additional genes encoding chemokine receptors.

With oGPCR DNAs in hand and with nearly all known ligands assigned, the task now is to use oGPCR DNAs to discover novel ligands[24]. The strategy employed is to express the oGPCR DNA in a cell and apply tissue extracts until a response is observed. The agonist ligand is then purified, synthesized and re-tested. This approach has been most successful in identifying neuropeptides. Peptide ligands often exhibit high-affinity interactions with their receptors, which enables detection at low concentrations and the development of radioligand binding assays. The first success at orphan ligand identification involved a GPCR with sequence identity to the opioid receptors. The natural ligand was identified by two research groups using brain extracts[2,3] and the peptide discovered was 17 amino acids in length, named either orphanin FQ or nociceptin. The peptide contains the tetrapeptide FGGF, which is similar to the motif YGGF of the opioid peptides. Another successful strategy used rat brain fractions that were applied to cells and Ca^{2+} mobilization measured; this succeeded in identifying a novel brain peptide. This peptide and a related peptide (from the same precursor protein) bound to two related oGPCRs and these peptides, which are found in the hypothalamus, function in appetite regulation and satiety control and thus were named orexins[25] (also known as hypocretins[26]). In a similar series f experiments, Hinuma *et al.*[27] measured arachidonate release from CHO cells transfected with the GPR10 (Ref. 28) to identify a novel brain peptide with prolactin-releasing properties at the anterior pituitary. This group has also identified another novel peptide, apelin[29], as the ligand for the receptor APJ (Ref. 30).

The elusive nature of certain labile natural agonists could be a significant hindrance to the discovery of oGPCR ligands, as there is no reason to believe that the remaining oGPCR ligands will all prove to be peptides. An attempt to address this problem involves the use of combinatorial chemistry to generate large libraries of compounds to be tested as surrogate agonists. Although not the physiological solution to the problem, such compounds are tools for probing the pharmacology of an oGPCR. Recently, an interesting variation to this approach was reported. Yeast expressing the human formyl peptide receptor-like oGPCR, FPR2 (Ref. 31), was made dependent on stimulation of this receptor for growth in histidine-free medium and then transfected with a plasmid DNA library designed to express random tridecapeptides. Yeast colonies that were no longer dependent on histidine were judged to have undergone autocrine stimulation and the responsible plasmids recovered. The results yielded a set of six peptides, one of which elicited Ca^{2+} mobilization in HEK293 cells transfected with the FPR2 plasmid.

Ligand-sceening assays

There has been a concerted effort to make ligand identification more efficient by developing cell-based assay systems that have low endogenous GPCR background or report G-protein activation events, or both, in a robust, readily detected manner. The existence of endogenous GPCR signalling systems is important because overexpression of one GPCR can elicit an exaggerated response via other, unrelated and previously unrecognized endogenous GPCRs (Ref. 32), and this could result in false positives. The aforementioned yeast expression system is attractive because of the absence of many endogenous GPCRs. In essence, it involves replacing the endogenous pheromone receptor with a mammalian GPCR and redirecting the pheromone pathway response from a mitogen-activated protein kinase type activation to a biosynthetic circuit, thus allowing the synthesis of histidine. In this case, agonist stimulation allows growth on histidine-free medium. Potential drawbacks of the yeast expression system are the difficulties in expressing some GPCRs achieving effective receptor–G-protein coupling and ligand binding to yeast cell wall components.

Another assay system, which uses mammalian cells, takes advantage of the relatively high expression levels achieved following transfection of oGPCR DNAs so that the endogenous, low-level receptors do not interfere. This system uses the translocation of β-arrestin to receptor sites on the plasma membrane after agonist-mediated receptor activation. Barak *et al.* have shown, using a β-arrestin-2/green fluorescent protein (βarr2–GFP) fusion protein and confocal microscopy, that on agonist stimulation of the β_2-adrenoceptor, βarr2–GFP translocates to the plasma membrane, and that this interaction can be enhanced by co-expression of G-protein-coupled receptor kinase 2 (Ref. 33). This group also showed that similar responses are observed with other receptors coupled to different G proteins, which suggests that the cellular visualization of the agonist-mediated translocation of βarr2–GFP cou provide a widely applicable method for detecting t activation of GPCRs.

A system that is useful in measuring GPCR-mediat activation of $G\alpha_q$, $G\alpha_{i/o}$ and $G\alpha_s$ is based on pigment d persion or aggregation in cultured *Xenopus laevis* melar phores[34,35]. Increases in cAMP ($G\alpha_s$-coupled receptors) activation of protein kinase C ($G\alpha_q$) lead to pigment d persion causing darkening of the cells, while decreases cAMP ($G\alpha_{i/o}$) lead to pigment aggregation near the r cleus and make the cells appear clear[36]. These colc changes are detected readily, however these cells hav substantial complement of endogenous GPCRs, whi could confound the results. Overexpression of recept in melanophores results in changes in the 'basal' signalli and promotes either the clear or the dark cell colour, th predicting either $G\alpha_{i/o}$ signalling or $G\alpha_q$ or $G\alpha_s$ pathwa

A simpler approach to detecting the activation of m tiple types of G proteins uses Gα16 as a universal adap G protein that can funnel the signal-transduction mach ery down a common pathway, such that a single secor messenger response (Ca^{2+} mobilization) can be measu for a given receptor[37]. Heterologous expression of Go allows the coupling of a wide range of GPCRs to phospl lipase activity, and thence to Ca^{2+} mobilization. For exa ple, the β_2-adrenoceptor normally couples only to G but when the β_2-adrenoceptor and Gα16 are transien co-expressed in COS7 cells agonist-dependent stin lation results in inositol phosphate (IP) productio Receptors linked to $G\alpha_s$ (e.g. dopamine D1, vasopressin and adenosine A_{2A} receptors) or pertussis-toxin-sensit $G\alpha_i$ (e.g. muscarinic acetylcholine M_2, 5-HT_{1A}, form peptide FPR1 and δ-opioid receptors), when co-transfec with Gα16, also caused concentration-dependent, agoni mediated IP generation[38]. Other receptors (e.g. thro boxane A_2 and vasopressin V_1) that routinely couple to C and Gα11 to stimulate IP generation were also shown couple effectively to Gα15 and Gα16 (Ref. 38). Howev this coupling is not universal, as the chemokine recept CCR1, that effectively couples to $G\alpha_i$ and $G\alpha_q$, failed couple to Gα16 (Ref. 39).

Other considerations

Recently, new complexities have been added to general approach to studying orphan GPCRs. For instar the oGPCR calcitonin receptor-like receptor, has b cloned[40]. The expression of this receptor was consist with the expression pattern of a calcitonin gene-rela peptide (CGRP). The efficient binding of CGRP or amy or both, to this receptor required the co-expression cofactor protein called receptor activity modifying pro 1 (RAMP1)[41].

Studies have shown that heterodimerization of GPCR subunits are required for the formation of a fu tional $GABA_B$ receptor[42–46]. The apparent requirement two different gene products to create a GPCR signal entity indicates that the characterization of some oGP might be more complex, perhaps indicating that functi

assays should begin to include co-expression of related oGPCRs.

In principle, the elimination of a GPCR gene from the germline and testing the resulting knockout mice for some change might provide clues to GPCR function, if not ligand identity. For example, when the mouse BLR1 orphan receptor was disrupted, it yielded mice with abnormal primary follicles and germinal centres of the spleen and Peyer's patches, reflecting the inability of B lymphocytes to migrate into B-cell areas[47]. A novel peptide that binds and activates BRL-1 was recently discovered from the EST database[48,49].

In view of the number of novel GPCRs that have been cloned and are continuing to be discovered, it is expected that many endogenous ligands will be discovered. Unquestionably, this will result in an increase in the knowledge of the diversity in intercellular signalling mechanisms and should lead to novel insights into complex or poorly understood human disorders; it will also expand the boundaries of pharmacology. In conclusion, the discovery of the endogenous ligands will help determine the precise physiological role for each oGPCR. As the functions of these novel receptors are uncovered, they could become targets for the development of new pharmacological therapies for diseases not previously considered amenable to pharmacological therapy.

Selected references

1 Marchese, A., George, S. and O'Dowd, B. (1998) in *Identification and Expression of G Protein-coupled Receptors* (Lynch, K., ed.), pp. 1–26, John Wylie & Sons
2 Civelli, O. *et al.* (1997) *J. Recept. Signal. Transduct. Res.* 17, 545–550
3 Mollereau, C. *et al.* (1999) *Mol. Pharmacol.* 55, 324–331
4 Zondag, G. *et al.* (1998) *Biochem. J.* 330, 605–609
5 Okamoto, H. *et al.* (1998) *J. Biol. Chem.* 273, 27104–27110
6 Lee, M. J. *et al.* (1998) *Science* 279, 1552–1555
7 An, S. *et al.* (1998) *J. Biol. Chem.* 273, 7906–7910
8 Hecht, J. H. *et al.* (1996) *J. Cell Biol.* 135, 1071–1083
9 An, S. *et al.* (1998) *Mol. Pharmacol.* 54, 881–888
10 Maenhaut, C. *et al.* (1990) *Biochem. Biophys. Res. Commun.* 173, 1169–1178
11 Matsuda, L. A. *et al.* (1990) *Nature* 346, 561–564
12 Yoshida, R. *et al.* (1997) *J. Biol. Chem.* 272, 13803–13809
13 Hieshima, K. *et al.* (1997) *J. Biol. Chem.* 272, 5846–5853
14 Power, C. A. *et al.* (1997) *J. Exp. Med.* 186, 825–835
15 Liao, F., Lee, H. and Farber, J. (1997) *Genomics* 40, 175–180
16 Baba, M. *et al.* (1997) *J. Biol. Chem.* 272, 14893–14898
17 Wells, T. N. and Peitsch, M. C. (1997) *J. Leukocyte Biol.* 61, 545–550
18 Legler, D. F. *et al.* (1998) *J. Exp. Med.* 187, 655–660
19 Pan, Y. *et al.* (1997) *Nature* 387, 611–617
20 Bazan, J. F. *et al.* (1997) *Nature* 385, 640–644
21 Imai, T. *et al.* (1997) *Cell* 91, 521–530
22 Heiber, M. *et al.* (1995) *DNA Cell Biol.* 14, 25–35
23 Yoshida, T. *et al.* (1998) *J. Biol. Chem.* 273, 16551–16554
24 Stadel, J. M., Wilson, S. and Bergsma, D. J. (1997) *Trends Pharmacol. Sci.* 18, 430–437
25 Sakurai, T. *et al.* (1998) *Cell* 92, 573–585
26 de Lecea, L. *et al.* (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 322–327
27 Hinuma, S. *et al.* (1998) *Nature* 393, 272–276
28 Marchese, A. *et al.* (1995) *Genomics* 29, 335–344
29 Tatemoto, K. *et al.* (1998) *Biochem. Biophys. Res. Commun.* 251, 471–476
30 O'Dowd, B. F. *et al.* (1993) *Gene* 136, 355–360
31 Klein, C. *et al.* (1998) *Nat. Biotechnol* 16, 1334–1337
32 Monnot, C. *et al.* (1991) *Mol. Endocrinol.* 5, 1477–1487
33 Barak, L. S., Ferguson, S. S. G., Zhang, J. and Caron, M. G. (1997) *J. Biol. Chem.* 272, 27497–27500
34 Graminski, G. F., Jayawickreme, C. K., Potenza, M. N. and Lerner, M. R. (1993) *J. Biol. Chem.* 268, 5957–5964
35 Potenza, M. N., Graminski, G. F. and Lerner, M. R. (1992) *Anal. Biochem.* 206, 315–322
36 McClintock, T. S. *et al.* (1993) *Anal. Biochem.* 209, 298–305
37 Milligan, G., Marshall, F. and Rees, S. (1996) *Trends Pharmacol. Sci.* 17, 235–237
38 Offermanns, S. and Simon, M. I. (1995) *J. Biol. Chem.* 270, 15175–15180
39 Arai, H. and Charo, I. F. (1996) *J. Biol. Chem.* 271, 21814–21819
40 Njuki, F. *et al.* (1993) *Clin. Sci.* 85, 385–388
41 McLatchie, L. *et al.* (1998) *Nature* 393, 333–339
42 Jones, K. A. *et al.* (1998) *Nature* 396, 674–679
43 White, J. H. *et al.* (1998) *Nature* 396, 679–682
44 Kaupmann, K. *et al.* (1998) *Nature* 396, 683–686
45 Ng, G. *et al.* (1999) *J. Biol. Chem.* 274, 7607–7610
46 Kuner, R. *et al.* (1999) *Science* 283, 74–77
47 Forster, R. *et al.* (1996) *Cell* 87, 1037–1047
48 Legler, D. F. *et al.* (1998) *J. Exp. Med.* 187, 655–660
49 Gunn, M. D. (1998) *Nature* 391, 799–803

Acknowledgements
The authors' research i supported by grants fro the Medical Research Council of Canada, the National Institute for Drug Abuse, and the Smokeless Tobacco Research Council.

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126
‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

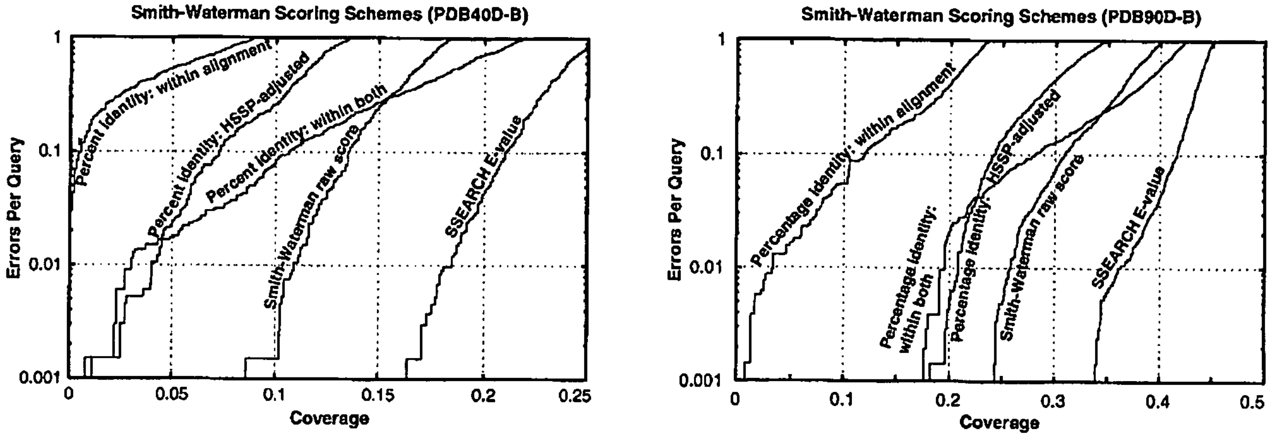

FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the *x* axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The *y* axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

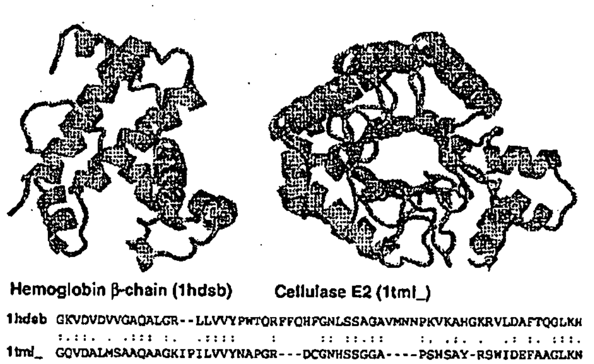

FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

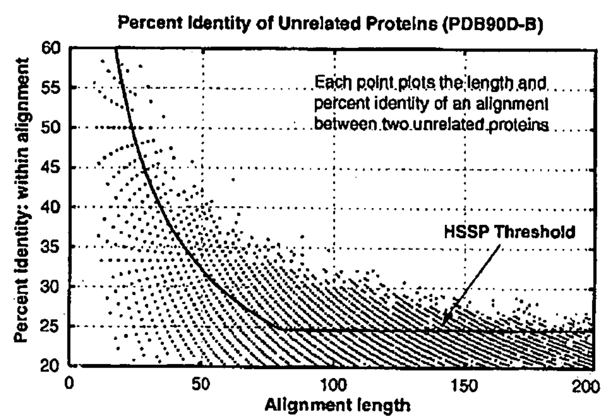

FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

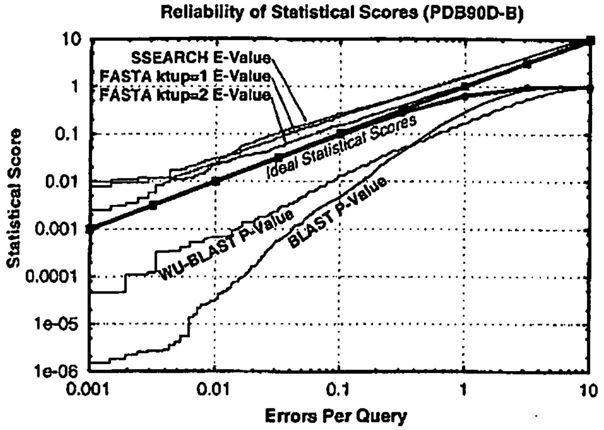

FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

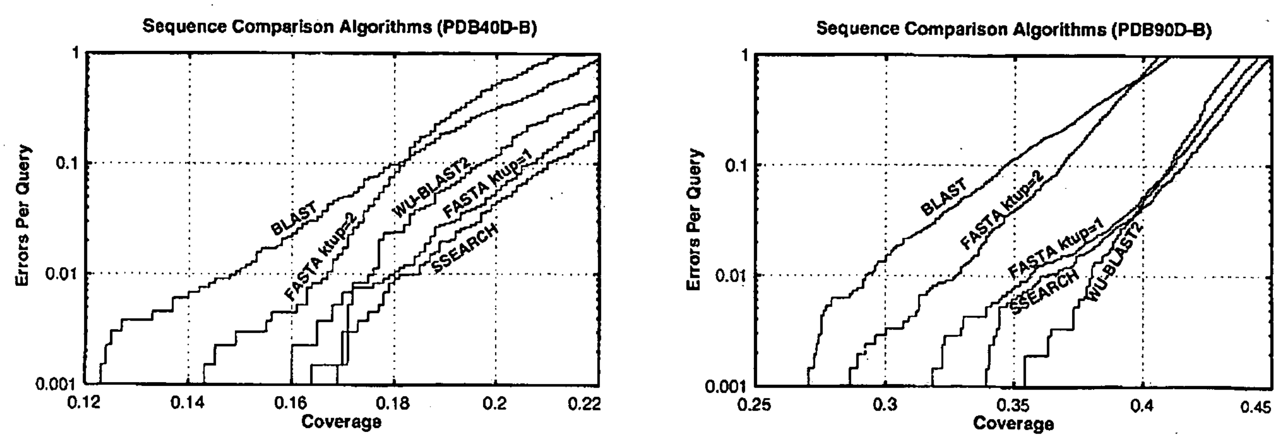

FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

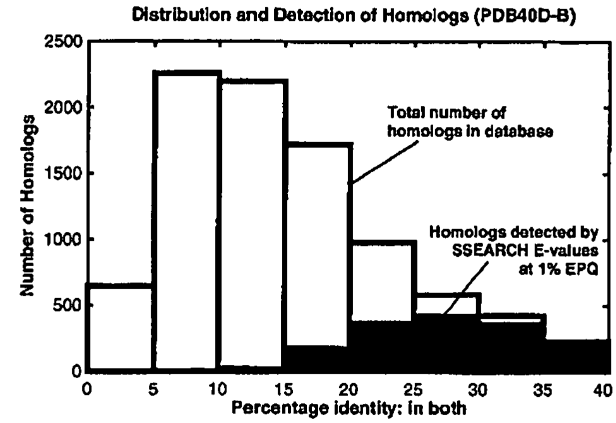

FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266,** 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266,** 635–643.
6. Pearson, W. R. (1991) *Genomics* **11,** 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4,** 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266,** 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249,** 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17,** 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24,** 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24,** 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233,** 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1,** 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1,** 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14,** 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12,** 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9,** 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7,** 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5,** 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39,** 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20,** 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4,** 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131,** 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32,** 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.